DATS 6101 Introduction to Data Science

Project 1 Outline (Spring 2020)

Goal: Better understand the initial stages of a data focused project by conducting background research and completing Exploratory Data Analysis (EDA).

I. Development of a **research driven question (SMART)** focused on a dataset either inside of R or one of your choosing from any online sources. Acceptable dataset for this "big data" class requires 3000+ observations (that is, 3000+ rows of data for the data frame).

II. Provide an **R-markdown file**, knitted into **HTML**, which shows the R-code and brief explanations as well as the rationale of the Exploratory Data Analysis of your project. (**Also** submit your **data file**, or give the online source url.) This document shows a technical person the math/stat/codes that you used in your analysis. It should include:
   - Summary of the dataset
   - Descriptive Statistics
   - Graphical representations of the data
   - [When applicable] Measures of Variance / sd
   - [When applicable] Normality tests
   - [When applicable] Initial correlation / Chi Square tests / ANOVA analysis / Z-test or Z-interval / T-test or T-interval etc.

III. Write a roughly 10-page (definitely no more than 4000 words, charts do not count) summary of the research and EDA process of your project. The summary should be prepared in **R-markdown**, and knitted into **HTML**. You may take some of the work in part II (such as graphs and results) to include here. They can overlap. This summary is to-be presented to your boss, your client, or to-be submitted for publication in journals. Potential area of topics to address in this summary may include:
   - What do we know about this dataset?
   - What are the limitations of the dataset?
   - How was the information gathered?
   - What analysis has already been completed related to the content in your dataset?
   - How did the research you gathered contribute to your question development?
   - What additional information would be beneficial?
   - How did your question change, if at all, after Exploratory Data Analysis?
   - Based on EDA can you begin to sketch out an answer to your question?
   - References (APA style preferred)

IV. Develop a **20-to-25-minute** presentation for the team that effectively communicates the results of these initial stages of a data science project to be presented during class.

Grading:

I. *
II. 25%
III. Together with part I, total of 50%
IV. 25% (Individually graded)

Grades for parts I through III are team-based. But I reserve the right to award different grades to team members if there is evidence of unfair contribution amount within the group.